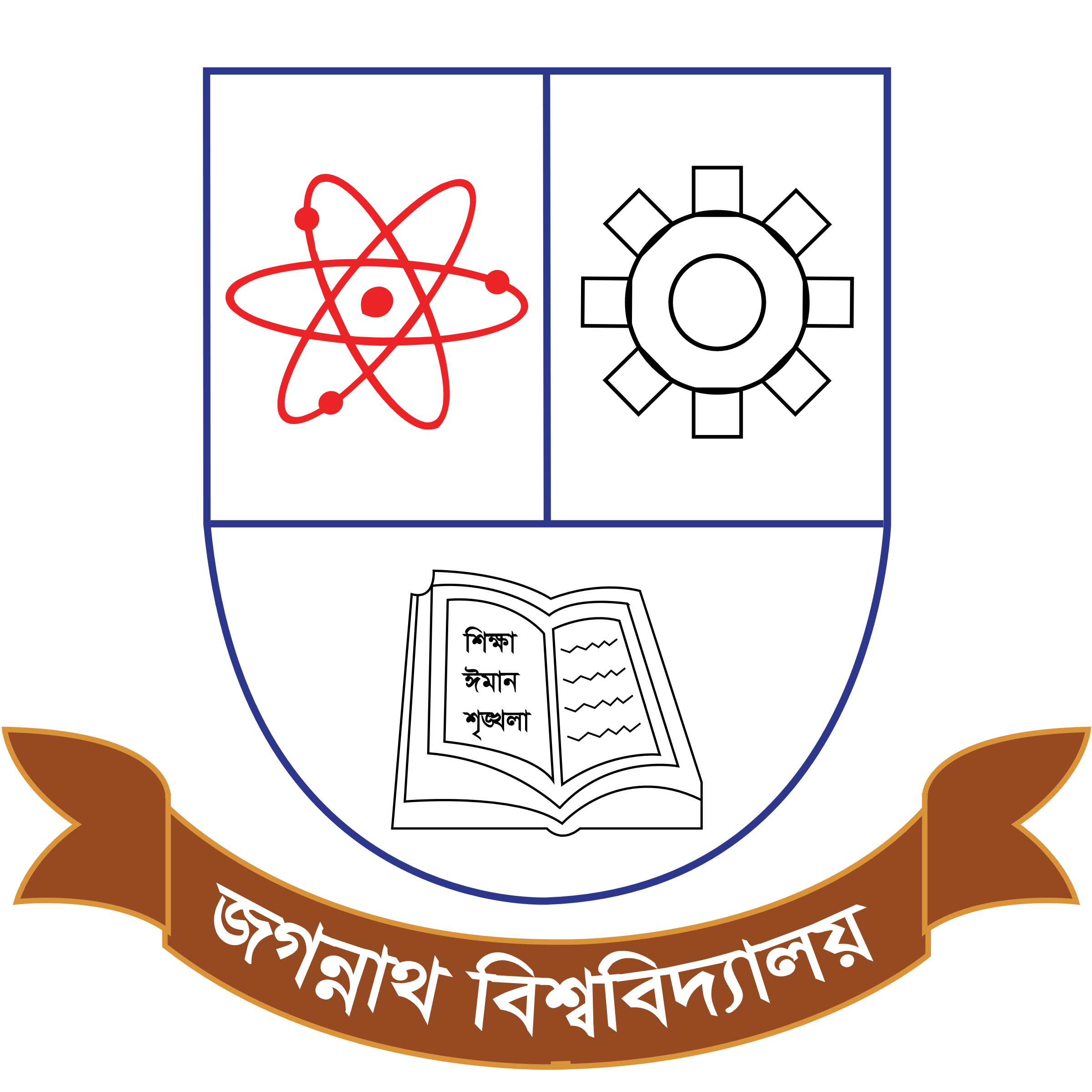
****

**Data Science Lab**

CSEL-42--

|  |
| --- |
| Assignment on Exploratory Data Analysis |

|  |  |  |
| --- | --- | --- |
| Submitted by | | |
|  |  |  |
| **NISHAT MAHMUD**  ID: B190305003 | **FAHIM HASAN**  ID: B190305029 | **MD. WALIUL ISLAM RAYHAN**  ID: B190305034 |

Submitted To

**MD. MANOWARUL ISLAM, PHD**

Associate Professor

Dept. of CSE

Jagannath University, Dhaka - 1100

Department of Computer Science and Engineering, Jagannath University, Dhaka

The Titanic dataset is a classic dataset widely used in data science and machine learning for exploring relationships between features and outcomes. It provides detailed information about passengers on the Titanic, including demographic details, travel class, and survival status.

The primary goal of this assignment is to perform Exploratory Data Analysis (EDA) to uncover meaningful insights about the dataset. In addition, data preprocessing steps are included to handle missing values and transform features to enable more effective analysis. Key visualizations such as histograms, bar plots, and heatmaps are used to highlight trends and correlations in the data.

This document provides a detailed explanation of each code section, ensuring a clear understanding of the EDA process and its findings.

A screenshot of a computer

Description automatically generated

This section imports all the necessary libraries for data processing, visualization, and modeling. %matplotlib inline ensures that plots are displayed directly within the Jupyter Notebook.

A white background with black dots

Description automatically generated

Loads the Titanic dataset from Seaborn’s built-in datasets.

A white background with black dots

Description automatically generated

Displays the first five rows of the dataset to preview its structure and contents.

A close-up of a computer screen

Description automatically generated

A white background with black dots

Description automatically generated

Provides a concise summary of the dataset, including column data types and the presence of null values.

A screenshot of a computer program

Description automatically generated

A close-up of a white background

Description automatically generated

Displays statistical summaries such as mean, median, and standard deviation for numerical columns.

A screenshot of a computer code

Description automatically generated

A white background with black dots

Description automatically generated

Counts the number of missing values in each column to identify issues with incomplete data.

A white background with black and white clouds

Description automatically generated with medium confidence

A close-up of a computer code

Description automatically generated

Fills missing values in the 'age' column with the median value using the SimpleImputer. Also, replaces missing values in 'embarked' and 'embark\_town' columns with the most frequently occurring value.

A close-up of a white background

Description automatically generated

Drops the 'deck' column because it contains too many missing values, making it unreliable for analysis.

A white background with black dots

Description automatically generated

Removes rows with missing values in the 'alive' column to maintain consistency.

A white background with blue and white symbols

Description automatically generated

Creates a new feature 'family\_size' by summing the number of siblings/spouses ('sibsp') and parents/children ('parch') on board, adding 1 for the passenger.

A close-up of a computer screen

Description automatically generated

Encodes the 'sex' column into numeric values (0 for female, 1 for male) using label encoding.

A green text on a white background

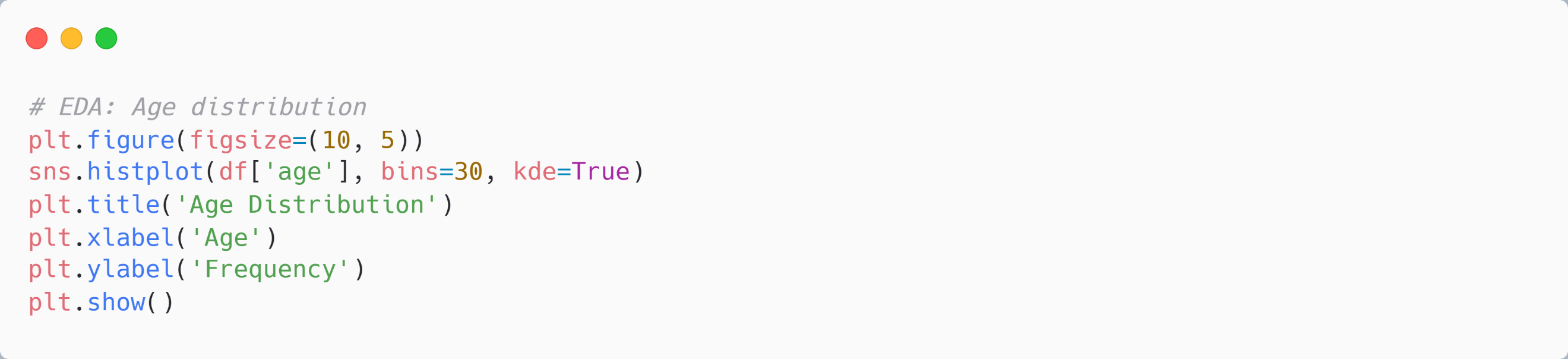
Description automatically generated

Performs one-hot encoding on categorical columns to convert them into numerical format, dropping the first category to avoid redundancy.

A close-up of a white background

Description automatically generated

Standardizes the 'age' and 'fare' columns to have a mean of 0 and a standard deviation of 1, improving model performance.



Plots the distribution of 'age' using a histogram with a kernel density estimate (KDE) curve to visualize the spread of ages in the dataset.

A graph of a number of age distribution

Description automatically generated

A close-up of a white background

Description automatically generated

Plots the distribution of 'fare' to examine its spread and detect potential outliers.

A graph of a number of bars

Description automatically generated with medium confidence

A white background with black and red text

Description automatically generated

Creates a bar plot to show survival rates for different genders.

A graph of survival rate by gender

Description automatically generated

A close-up of a computer screen

Description automatically generated

Visualizes survival rates by passenger class to observe how class influenced survival.

A graph of a graph showing a number of blue squares

Description automatically generated with medium confidence

A close-up of a computer screen

Description automatically generated

Displays a heatmap to visualize the correlation between features in the dataset.

A screenshot of a computer

Description automatically generated

A close-up of a computer screen

Description automatically generated

Uses a box plot to detect outliers in the 'fare' column, categorized by survival status.

A graph of a number of people

Description automatically generated with medium confidence

A close-up of a white background

Description automatically generated

Separates features (X) and target variable (y) for logistic regression.

A close-up of a white background

Description automatically generated

Splits the dataset into training (70%) and testing (30%) sets for model evaluation.

A white background with black and white clouds

Description automatically generated with medium confidence

Initializes and trains a logistic regression model with a maximum of 1000 iterations for convergence.

A close-up of a blue and white sign

Description automatically generated

A white background with black dots

Description automatically generated

Uses the trained model to predict survival outcomes for the test set.

A close-up of a white screen

Description automatically generated

Evaluates model performance using accuracy, a classification report, and a confusion matrix.

A screenshot of a computer

Description automatically generated

Through the Exploratory Data Analysis (EDA) of the Titanic dataset, we uncovered valuable insights about survival rates and their relationship with features such as age, gender, fare, and passenger class.

The analysis also showcased the importance of data preprocessing, including handling missing values and feature engineering, to enable accurate and meaningful exploration. These insights can serve as a foundation for further predictive modeling or decision-making.